置信区间宽度等高线图在线性混合效应模型样本量规划中的应用*

刘　玥[1]　徐　雷[1] 刘红云[2,3] 韩雨婷[4] 游晓锋[5] 万志林[1]

([1]四川师范大学脑与心理科学研究院, 成都 610066)

([2]应用实验心理北京市重点实验室) ([3]北京师范大学心理学部, 北京 100875)([4]北京语言大学，心理学院，北京，100083)([5]南昌师范学院数学与信息科学学院，南昌 360111)

摘要 线性混合效应模型在分析具有嵌套结构的心理学实验数据时具有明显优势。本文提出了置信区间宽度等高线图用于该模型的样本量规划。通过等高线图，确定同时符合检验力、效应量准确性以及置信区间宽度要求的被试量和试次数。结合关注被试内实验效应和被试变量调节效应的两类典型模型，通过两个模拟研究，采用基于蒙特卡洛模拟方法，探索效应量、随机效应大小和被试变量类型对置信区间宽度等高线图及样本量规划结果的影响。

关键词 线性混合效应模型，多水平模型，检验力分析，效应量，置信区间宽度

1 引言

近年来，心理学研究者对学术不端和研究可重复性问题的讨论日趋激烈。国内外越来越多的学术期刊推行预注册(pre-register)制度，能够有效避免根据结果决定是否继续收集数据的不良行为(例如，*p*-hacking)，促进科研过程和结果的公开透明，提高研究的可重复性(Nosek et al., 2022)。预注册时，对被试量、试次数等与研究设计相关的要素需有明确规划和充分理由。如何针对特定的统计模型开展样本量规划，是心理学研究者关心的问题。本研究基于线性混合效应模型，探索使用模拟方法结合检验力和效应量准确性开展样本量规划的范式，并通过开发直观的置信区间宽度等高线图，方便应用研究者确定符合要求的被试量和试次数，为开展研究设计、保证研究质量提供方法支持。

1.1 线性混合效应模型的样本量规划问题

随着研究问题的深入和数据收集手段的进步，含有随机效应的刺激和嵌套结构的设计越来越普遍。例如，心理语言学实验研究通常会使用词语作为刺激，但不同词语诱发的反应速度不同，会造成观察到的实验效应有一部分是由不同的词语刺激引起的(Barr et al., 2013)。此时，以传统方差分析为代表的方法由于混淆了实验效应与随机效应，会导致第 I 类错误和

* 收稿日期：2023-01-04

国家自然科学基金项目(32071091, 32200920)。

通信作者: 刘红云, E-mail: hyliu@bnu.edu.cn

检验力的估计偏差(Barr et al., 2013; Judd et al., 2017)。线性混合效应模型(Linear Mixed-Effects Models，LMEMs)可以避免由于对被试接受的同一条件下所有刺激求均值等方式(如，重复测量方差分析)造成的信息损失，且同时灵活考虑不同原因(如，刺激随机取样、被试嵌套结构等)造成的随机效应。因此，LMEMs 在心理学实验中的应用越来越广泛(Barr et al., 2013; Brauer & Curtin, 2018; Judd et al., 2017; Lee, 2018)。在 web of science 中检索近五年的心理学实验类论文，使用 LMEMs 约是使用方差分析的 1.5 倍。

然而，目前国内 LMEMs 的应用还很少。例如，2020-2022 年我国心理学顶刊《心理学报》上发表的 181 篇实验类文章中，仅 9 篇使用了 LMEMs，且其中的 5 篇没有阐述确定样本量的理由，3 篇应用 G * power 近似得到所需样本量，仅有 1 篇应用 simr 软件包采用模拟方法基于检验力分析确定样本量。制约该模型广泛应用的一个重要原因是，设计中随机效应的增加带来了模型复杂程度的增加，导致常用的样本量规划软件(例如 G * power 等)不再适用，研究者对基于 LMEMs 如何科学地规划实验设计，设置合理的被试量和试次数感到无所适从，急需方便易用的程序或图示，指导样本量规划。

1.2 基于检验力分析规划样本量

传统样本量规划主要基于虚无假设显著性检验(Null Hypothesis Significance Test, NHST)的检验力分析，要求样本量必须使检验力达到预设标准。检验力分析可分为公式推导方法和基于蒙特卡洛模拟方法(例如，Arend & Schäfer, 2019)。公式推导方法含有关于分布的强假设，当数据不符合时可能得到有偏差的结果(Judd et al., 2017)。基于蒙特卡洛模拟的方法是在预设的参数下基于特定模型重复生成数据，再基于模拟数据估计参数，统计所有重复中得到显著性结果的比例。其优势在于不需要推导参数分布，能够处理非正态分布的数据，并且可以灵活定义模型。一些学者已经开发了成熟的 R 软件包(如 simr)应用蒙特卡洛模拟的方式计算 LMEMs 的检验力(Green & MacLeod, 2016)。

为了方便应用研究者基于检验力分析确定适用于嵌套数据分析的合适样本量，一些研究者在模拟方法的基础上，开发了直观的图示以及配套程序，展示不同样本量情况下的检验力，为样本量规划提供参考。应用最广的是以样本量为横坐标，检验力为纵坐标的折线图(例如, Kumle et al., 2021)。研究者根据预设检验力做出水平线，与折线交点所对应的横坐标就是满足要求的最小样本量。Murayama 等(2022)还开发了生成检验力折线图的在线程序。但是，嵌套结构的数据需要确定两个水平样本量，不同实验设计下增加不同水平样本量的成本不同。折线图仅能固定某个水平样本量，以另一个水平样本量为横坐标生成，无法同时呈现

两个水平样本量与检验力的关系。Schultzberg 和 Muthén(2018)将水平 1、2 样本量分别作为横、纵坐标，用阴影区域表示符合检验力要求的两个水平样本量组合范围。Baker 等(2021)提出了检验力等高线图，将相同检验力的两个水平样本量组合的点连成等高线，用多条等高线表示不同检验力水平。综上，对于嵌套数据，研究者需要在同一个图内观察到两个水平样本量在检验力上的补偿关系，并在考虑实验成本的基础上综合权衡，得到合适的各水平样本量。

1.3 基于效应量准确性分析规划样本量

以上总结的样本量规划图示仅考虑了检验力。但是，随着学术界对 NHST 的批判，美国统计协会发表了关于谨慎使用 NHST 的声明，强调应避免仅报告显著性，而应同时报告效应量(Wasserstein & Lazar, 2016)及其区间估计的结果。因此，一些学者提出应基于效应量准确性分析开展样本量规划。

效应量准确性分析的核心是控制效应量置信区间(Confidence Interval，CI)的宽度，越窄表明其估计越准确(Maxwell et al., 2008)。有研究根据期望的 CI 上下限，倒推可接受的最大 CI 宽度(Usami, 2020)。例如，在效应量的点估计值为 0.5 的情况下，计算得到其 95%置信区间(以下简称“95% CI”)宽度为 0.6，则 95%CI 约为[0.2,0.8]。根据 Cohen(2013)的标准，该区间涵盖了效应量小、中、大的条件(0.2,0.5,0.8)，估计精确性差(Maxwell et al., 2008; Usami, 2020)。有的研究直接根据不同 CI 宽度计算对应的最小样本量(例如, Kelley & Rausch, 2006)。总之，目前关于如何确定可接受的最宽 CI 宽度仍未形成一致结论(例如, Kelley et al., 2018)。

为了方便应用研究者基于效应量准确性分析确定适用于嵌套数据分析的样本量，Hecht 和 Zitzmann(2021)提出了基于被试数和时间点的总体表现图，分别以二者作为横、纵坐标，通过收敛比例，参数估计偏差等指标计算模型拟合的综合表现得分，并以色块区分不同得分。研究者可以根据色块，权衡得到合适的样本量组合。但该图并未考虑检验力，并且色块仅表示综合得分，具有一定的主观性，研究者无法从图中清晰了解所关心的参数估计的准确性。

1.4 问题提出

综上，针对嵌套数据的样本量规划需同时保证检验力和效应量准确性达到要求。然而，已有的方法、程序或图示大多只基于其中一个目的展开 (例如, Arend & Schäfer, 2019; Kumle et al., 2021; Usami, 2020)，尚没有图示能够方便研究者同时考虑两方面要求规划样本量。因此，本研究提出 CI 宽度等高线图，采用蒙特卡洛模拟方法进行检验力和效应量准确

性分析，在图中同时呈现两个水平样本量不同组合下的检验力和 CI 宽度情况。由于 CI 宽度尚没有统一标准，本研究结合已有研究的两种思路，提供不同 CI 宽度下的样本量，建议研究者结合期望的 CI 上下限推出可接受的最宽 CI 宽度，进而综合检验力分析结果确定被试量和试次数的理想结合点。

此外，在以心理学实验研究为背景的样本量规划中，研究者普遍关注基于实验效应中固定效应的样本量规划(Lee, 2018)，而不关注基于被试变量对实验效应的调节效应的样本量规划。然而，随着心理学个体差异视角研究的深入，越来越多的研究开始探索不同类型个体间的实验效应是否存在差异。例如，蒋元萍等(2022)发现，积极情绪和消极情绪状态下被试(被试调节变量)的跨期决策行为(实验效应)存在显著差异。这类研究需要样本量规划满足被试变量调节效应估计准确性的要求。因此，本研究以典型的被试内重复实验设计为背景，基于 LMEMs，分别探讨基于被试内变量的实验效应和被试间变量的调节效应的样本量规划问题。

本文首先在多层线性模型框架下重构模型，以更好适应实验设计在不同层级加入自变量(控制变量)的需要。然后，说明生成 CI 宽度等高线图的流程及其函数。最后，分别基于被试内变量的实验效应和被试间变量的调节效应进行模拟研究，考察实验效应、随机斜率、被试变量类型如何影响评价指标结果和 CI 宽度等高线图，并说明如何根据结果推荐合适的样本量。

2 心理学实验研究中的线性混合效应模型

LMEMs 的一般形式可见 Williams 等 (2021)的文章。在多层线性模型的框架下，可对其重新定义。以刺激嵌套于实验条件的被试内实验设计为例，假设刺激没有重复(Barr et al., 2013; Lee, 2018)。水平 1 表示试次(trial)水平，水平 2 表示被试水平，试次嵌套于被试。随机斜率模型(模型 1)可表示为

水平 1：$Y_{ji} = \beta_{0i} + \beta_{1i} X_{ji} + I_{0j} + r_{ji}$， (1)

水平 2：$= \gamma_{00} + u_{0i}$， (2)

$\beta_{1i} = \gamma_{10} + u_{1i}$， (3)

其中，Y_{ji} 表示连续的结果变量(j=1,…,J 表示试次，i=1,…,I 表示被试)，X_{ji} 表示实验效应的虚无编码，β_{0i} 和β_{1i} 分别表示被试的随机截距和随机斜率，即不

同被试基线水平和实验效应的不同，I_{0j} 表示刺激的随机截距(不同刺激的效应不同)。γ_{00}和γ_{10}分别表示被试随机截距的均值和随机斜率的均值，其中γ_{10}是实验效应的固定部分，是重点考察的效应量指标。u_{0i}，u_{1i}，r_{ji} 分别表示水平 2 截距、斜率的随机部分和水平 1 的残差。模型假设$r_{ij} \sim N(0,\sigma^2)$，$\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N(0,\boldsymbol{\Sigma})$，$\boldsymbol{\Sigma} = \begin{bmatrix} \tau_{00}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & \tau_{11}^2 \end{bmatrix}$，刺激的随机截距$I_{0j} \sim N(0,\omega_{00}^2)$。

多层线性模型的优势在于能够方便地在不同水平加入解释变量。例如，可在水平 2 加入自变量W_i，用于解释随机截距和随机斜率存在个体间差异的原因(模型 2)。

水平 1：
$$Y_{ji}\ \beta_{0i} + \beta_{1i}X_{ji} + I_{0j} + r_{ji}, \tag{4}$$

水平 2：
$$\gamma_{00} + \gamma_{01}W_i + u_{0i}, \tag{5}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + u_{1i}, \tag{6}$$

其中，W_i 表示被试变量，γ_{01}表示被试变量对随机截距的影响，γ_{11}表示被试变量对随机斜率的影响，也可看作水平 1 和水平 2 变量的跨水平交互作用，是重点考察的效应量指标。

3 置信区间宽度等高线图生成步骤

基于模拟的方法生成置信区间宽度等高线图实现样本量规划包含以下步骤。

第一，设置参数。在实验研究背景下，选用特定的 LMEM，设置水平 1、水平 2 样本量[1]，固定效应取值，以及随机效应分布。

第二，生成数据。基于步骤一中定义的模型重复生成数据 *N* 次(如，*N*=1000)。

第三，参数估计。对于每次重复，使用产生模型与数据拟合。应用 R 软件包 *lme4*(Bates et al., 2011)基于限制性极大似然(restricted maximum likelihood, REML)方法估计参数。采用默认的 Wald 方法计算效应量参数的 CI。

第四，变化水平 1、水平 2 样本量，重复步骤一到三。

第五，计算评价指标。详见 4.2。

[1] 当水平 1、水平 2 自变量为分类变量时，可设定不同类别的样本量。

第六，根据标准对评价指标作出判断，画出 CI 宽度等高线图，推荐合适的样本量。本研究建议采用效应量标准的最高水平减去最低水平作为可接受的最大 CI 宽度。

本研究基于 R 语言(R Development Core Team, 2019)编写了适用于 LMEMs 样本量规划的函数 samplesize_LMEM.R(见在线补充材料 2)。调用函数，并输入相应的参数运行程序，可以得到评价指标结果和 CI 宽度等高线图。应用流程如图 1 所示。调用语句及其说明请参考在线补充材料 3。本函数具有一定的灵活性，例如设置$\omega_{00}^{2} = 0$时，数据生成模型简化为不含刺激随机效应的模型；设置$\omega_{00}^{2} = 0, \tau_{11}^{2} = 0$时，简化为随机截距模型；设置$\omega_{00}^{2} = 0, \tau_{11}^{2} = 0, ICC = 0$时，简化为一般线性模型。

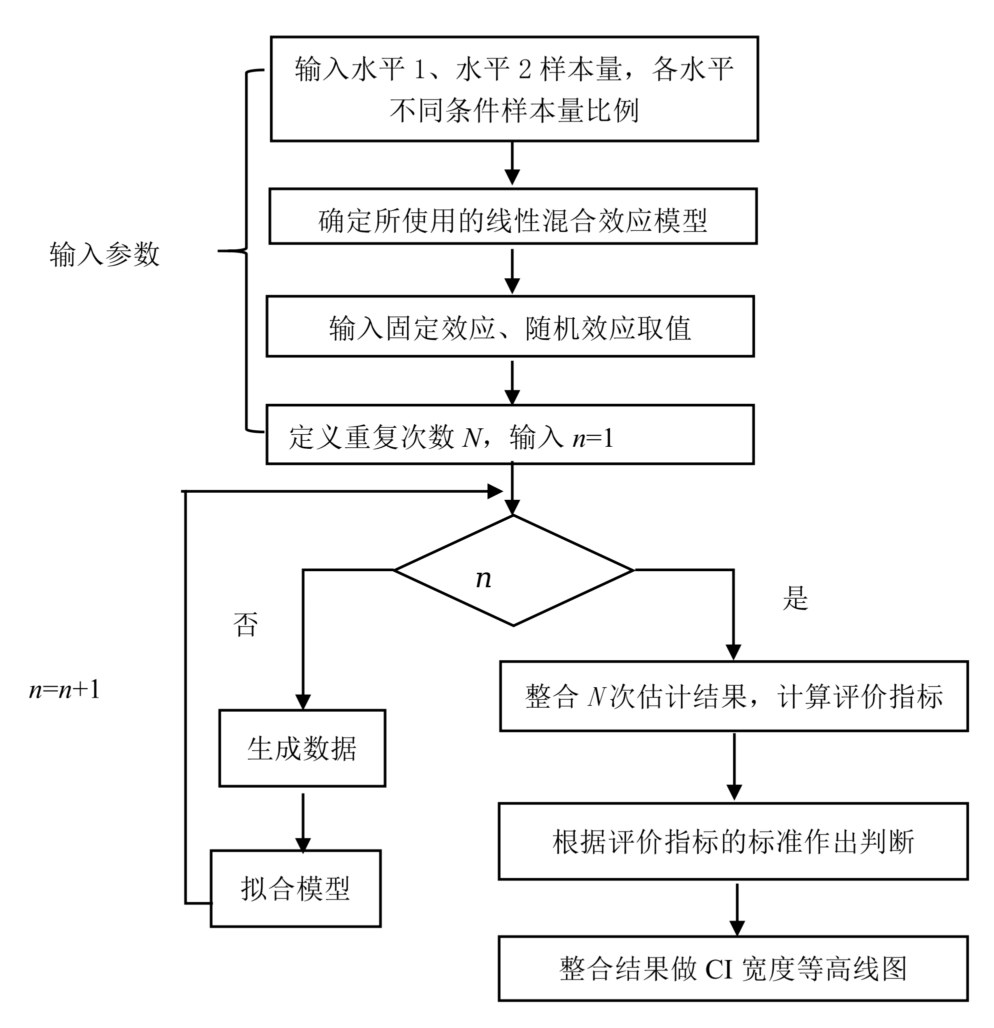


图 1 CI 宽度等高线图生成流程图

下面通过两个模拟研究，考察不同因素对检验力和效应量估计准确性的影响，说明 CI(本

研究为 95%CI)宽度等高线图在样本量规划中的应用。

4 模拟研究一：基于被试内实验效应的样本量规划

研究一在模型 1 的框架下，针对实验效应γ_{10}，即水平 1 自变量的固定效应，考察γ_{10}对模型估计结果的影响，并通过 CI 宽度等高线图提供样本量建议。

4.1 参数设置

4.1.1 固定参数设置

基于模型 1 模拟生成数据。参照 Arend 和 Schäfer(2019)的参数设置，随机截距的固定效应γ_{00}固定为 0，残差$r_{ij} \sim N(0,1)$。预研究发现，组内相关[2](intraclass correlation coefficient, ICC)大小对γ_{10}的检验力和参数估计准确性都没有显著影响，因此固定为中等水平 0.3（Arend & Schäfer, 2019），已知残差方差$\sigma^2 = 1$，根据下式，计算得到τ_{00}^2的值。

$$\tau_{00}^2 = ICC/(\sigma^2 - ICC). \tag{7}$$

标准化的随机斜率方差[3]固定为中等水平($\tau_{11.std}^2$=0.09)。为简化研究，随机截距和随机斜率的协方差固定为 0($\rho = 0$，Arend & Schäfer, 2019)。刺激的随机效应[4]固定为较小水平ω_{00}^2=0.2(Cho et. al., 2017) 。最后，根据残差方差，得到用于产生数据的总体模型的随机斜率方差。

$$\tau_{11} = \tau_{11.std} \times \sigma^2. \tag{8}$$

X_{ji}设定为二分类变量(如，控制组和实验组)。采用偏差编码(deviation coding, Barr et al., 2013; Lee, 2018)的形式，编码为-0.5 和 0.5。每种条件下重复模拟 1000 次(例如, Zhang, 2014)。

[2] 在多水平模型中，组内相关 ICC 用于表示零模型（不含任何预测变量的模型）中水平 2 变异占总变异的比例，值越大组间变异越大。一般而言，被试嵌套于组的设计所得到的 ICC 要小于测量嵌套于被试的 ICC。

[3] 预研究发现，标准化的随机斜率方差$\tau_{11.std}^2$对γ_{10}的检验力和参数估计偏差影响不大。

[4] 本研究暂不考察刺激随机效应大小对样本量规划结果的影响，因此参考 Cho 等（2017）的实证调查，将刺激的随机效应固定为较小水平 0.2。

4.1.2 变化参数设置

参考 Arend 和 Schäfer(2019)的研究，实验效应的大小($\gamma_{10.std}$[5])设为 3 个水平：0.2(小)、 0.5(中)、0.8(大)。在每种条件下分别进行样本量规划。

水平 1 样本量(J，试次数)，包含 10 个水平：10，20，30， 50，70，100，150，200，250，300。水平 2 样本量(I，被试量)，包含 9 个水平：10，30，50，70，100，200，400，600，800。共形成 $10 \times 9 = 90$ 种样本量组合[6]。

此外，有研究证明，当不同条件下试次数不等时(非平衡设计)，同等样本量条件下的检验力较小(Kumle et al., 2021)。因此，为考察非平衡设计对样本量规划的影响，在效应量中等的水平下，增加自变量两个类别样本量不等的情况。参考 Kumle 等(2021)的研究，设两个水平的样本量比例为 1:4。

综上，完成参数设置，调用 samplesize_LMEM.R 函数运行得到结果。

4.2 评价指标

评价指标包括 5 个方面。(1)**收敛率**。即参数估计收敛次数占总重复次数的比例。是否收敛采用 *lme4* 默认的 Hessian 检验评价(Bates et al., 2011)。后面的所有评价指标均基于收敛的情况计算。(2)**检验力**。γ_{10}的 CI 不包括 0 的次数占所有收敛次数的比例。预设的检验力标准为大于等于 0.8。(3)**效应量(固定效应)估计的准确性**。包括估计偏差(bias)，相对估计偏差(relative parameter estimation bias, rbias)，误差均方根(root mean squared error, RMSE)，CI 宽度(width)，CI 对真值的覆盖率(CP)。以γ_{10}为例：

$$bias = \frac{1}{\sum_{n=1}^{N} H^{(n)}} \sum_{n=1}^{N} H^{(n)} \left(\widehat{\gamma}_{10}^{(n)} - \gamma_{10}\right), \quad (9)$$

$$rbias = \frac{\left| \frac{1}{\sum_{n=1}^{N} H^{(n)}} \sum_{n=1}^{N} H^{(n)} \left(\widehat{\gamma}_{10}^{(n)} - \gamma_{10}\right) \right|}{\gamma_{10}}, \quad (10)$$

---

[5] 在多水平模型中，$\gamma_{10.std} = \gamma_{10} * SD_{predictor} / SD_{outcome}$。当自变量为分类变量时，$\gamma_{10.std}$ 为部分标准化的回归系数，即只对因变量标准化($SD_{outcome} = \sigma$，$\gamma_{10.std} = \gamma_{10}/\sigma$)。该系数代表了自变量两个类别在因变量上的标准化均值差异（Cohen's *d*）。

[6] 水平 1 样本量中，J=10 的水平代表了 Lee（2018）的研究中使用 Laplace 接近方法没有收敛问题的条件，J=300 的水平代表了 Schultzberg 和 Muthén（2018）关于动态结构方程模型样本量规划研究中测试时间点设置的最大水平。水平 2 样本量中，I=10 的水平接近 Lee（2018）总结的类似实验设计所使用的被试量最小值（16），I=800 的水平接近 Lee（2018）模拟研究中设置的 1000 名被试的水平，目的是为了探索大样本条件对效应量估计准确性提高的作用。最小到最大样本量水平之间的变化参考了同类样本量规划研究（例如, Schultzberg & Muthén, 2018）。

$$\mathrm{RMSE}=\sqrt{\frac{1}{\sum_{n=1}^{N}H^{(n)}}\sum_{n=1}^{N}H^{(n)}\left(\widehat{\gamma}_{10}^{(n)}-\gamma_{10}\right)^{2}},\tag{11}$$

$$\mathrm{width}=\frac{1}{\sum_{n=1}^{N}H^{(n)}}\sum_{n=1}^{N}H^{(n)}\times width^{(n)},\tag{12}$$

$$\mathrm{CP}=\frac{1}{\sum_{n=1}^{N}H^{(n)}}\sum_{n=1}^{N}H^{(n)}\times coverage^{(n)},\tag{13}$$

其中，γ_{10}表示真值，N 表示模拟重复次数。对于第 n 次重复，$\widehat{\gamma}_{10}^{(n)}$为$\gamma_{10}$估计值，$H^{(n)}$为估计结果是否收敛的指标变量，$H^{(n)}$=0 表示不收敛，$H^{(n)}$=1 表示收敛。$width^{(n)}$表示$\widehat{\gamma}_{10}^{(n)}$的 CI 宽度，$coverage^{(n)}$为$\widehat{\gamma}_{10}^{(n)}$的 CI 是否覆盖真值$\gamma_{10}$的指标变量，$coverage^{(n)}=0$表示没有覆盖真值，$coverage^{(n)}=1$表示覆盖真值。如果效应量$\gamma_{10}$估计准确，则 bias 应在 0 附近，rbias 应小于其临界值 0.1(Koch et al., 2014)，RMSE 应较小，width 应较窄，CP 应在 0.925 到 0.975 之间(Bradley, 1978)。(4)**效应量标准误估计的准确性**。为评价效应量标准误估计的准确性，计算了效应量的估计标准误相对于其估计值标准差的偏差(SE-SD bias)。以γ_{10}为例，

$$\mathrm{SE-SD\ bias}=\frac{1}{\sum_{n=1}^{N}H^{(n)}}\sum_{n=1}^{N}H^{(n)}\left(SE_{\widehat{\gamma}_{10}}^{(n)}-SD_{\widehat{\gamma}_{10}}\right),\tag{14}$$

其中，$SE_{\widehat{\gamma}_{10}}^{(n)}$表示第 n 次重复$\widehat{\gamma}_{10}^{(n)}$的估计标准误，$SD_{\widehat{\gamma}_{10}}$表示所有收敛的重复中$\widehat{\gamma}_{10}^{(n)}$的标准差。如果$\gamma_{10}$的估计标准误准确，则 SE-SD bias 应接近于0(Schultzberg & Muthén, 2018)。(5)**随机效应估计的准确性**。随机效应方差估计值(包括σ^2，τ_{00}^2，τ_{11}^2和ω_{00}^2)的 rbias。其计算方法与公式(10)类似。

4.3 研究结果

4.3.1 收敛情况

附表 1 和 2(在线补充材料 1)分别呈现了平衡和非平衡样本量分配条件下，随机斜率模型(模型 1)的收敛率。各条件下基本不存在收敛问题，收敛率均在 0.7 以上，两个水平样本量均小于 200 时，收敛率均超过 0.9。另外，**效应量大小和是否为平衡设计对收敛率几乎没**

有影响。

4.3.2 检验力结果

平衡设计各条件下检验力结果如表 1 所示。从表中可以看出，**效应量越大，检验力越大，满足 0.8 标准需要的样本量越小。**例如，被试量为中等水平(200 人)，当效应量为 0.2 时，需要 200 个试次才能保证检验力达到 0.8 及以上；而当效应量为 0.8 时，只需要 20 个试次就能保证检验力达到 0.8 及以上。非平衡设计的检验力结果见附表 3(在线补充材料 1)。对比发现，**非平衡设计的检验力普遍小于平衡设计的结果。**例如，当被试量为 10 人，检验力达到 0.8 时，平衡设计下需要 50 个试次，而非平衡设计下则需要 100 个试次。

表 1 研究一平衡设计各条件下线性混合效应模型水平 1 自变量效应的检验力

ES	*I*	*J*									
		10	20	30	50	70	100	150	200	250	300
0.2	10	0.107	0.112	0.131	0.168	0.181	0.224	0.279	0.312	0.369	0.379
	30	0.118	0.152	0.202	0.266	0.335	0.446	0.585	0.677	0.738	**0.802**
	50	0.170	0.175	0.224	0.278	0.409	0.490	0.677	0.756	**0.832**	**0.888**
	70	0.125	0.171	0.218	0.311	0.412	0.543	0.683	0.791	**0.866**	**0.930**
	100	0.133	0.169	0.233	0.335	0.420	0.535	0.701	**0.816**	**0.893**	**0.935**
	200	0.147	0.188	0.234	0.344	0.455	0.586	0.745	**0.845**	**0.913**	**0.951**
	400	0.115	0.194	0.232	0.345	0.433	0.574	0.766	**0.852**	**0.918**	**0.958**
	600	0.123	0.193	0.236	0.376	0.447	0.606	0.740	**0.878**	**0.931**	**0.965**
	800	0.147	0.202	0.245	0.377	0.480	0.549	0.764	**0.909**	**0.948**	**0.969**
0.5	10	0.298	0.481	0.626	**0.804**	**0.891**	**0.975**	**0.994**	**0.999**	**1.000**	**1.000**
	30	0.383	0.631	0.782	**0.927**	**0.986**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**
	50	0.438	0.659	**0.810**	**0.959**	**0.992**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	70	0.430	0.651	**0.822**	**0.963**	**0.992**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**
	100	0.453	0.659	**0.845**	**0.967**	**0.996**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**
	200	0.451	0.679	**0.846**	**0.968**	**0.999**	**0.998**	**1.000**	**1.000**	**1.000**	**1.000**
	400	0.453	0.714	**0.856**	**0.976**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	600	0.416	0.695	**0.849**	**0.972**	**0.994**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	800	0.464	0.715	**0.850**	**0.972**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
0.8	10	0.626	**0.876**	**0.959**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	30	0.715	**0.952**	**0.995**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	50	0.747	**0.956**	**0.993**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	70	0.773	**0.958**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	100	0.766	**0.968**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	200	0.766	**0.977**	**0.995**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	400	0.799	**0.970**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	600	0.783	**0.976**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	800	**0.805**	**0.973**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

注：*J* 表示水平 1 样本量，*I* 表示水平 2 样本量，*ES* 表示水平 1 自变量的效应量。表中加粗的为检验力大于等于 0.8 的结果。

4.3.3 效应量及其标准误估计准确性结果

效应量大小对效应量及其标准误估计准确性结果没有显著影响。表 2 呈现了平衡设计效应量为 0.5(中等)情况下效应量及其标准误估计准确性结果(只呈现 rbias，width 和 SE-SD bias 的结果，其他评价指标结果见附表 4，效应量为 0.2 和 0.8 的结果见附表 5、6，在线补充材料 1)。表 2 结果显示所有条件下 rbias 都小于 0.1。此外，附表 4 显示在所有条件下，bias 都在 0 附近波动；RMSE 较小，基本在 0.3 以下，且随着水平 1 和水平 2 样本量增加，尤其是水平 1 样本量增加，RMSE 减小；最后，除了水平 1 样本量为 10 的条件外，其他条件下的覆盖率都大于 0.925。以上结果说明各条件下，水平 1 自变量的固定效应估计准确。

根据效应量小和大的标准值 0.2 和 0.8，定义可接受的最宽 95%CI 宽度为 0.8-0.2=0.6。从表 3 看出，当水平 1 样本量为 30 及以下时，95%CI 宽度均超过了 0.6。说明在这些情况下效应量估计的标准误较大，导致其 95%CI 较宽。

最后，各种条件下 SE-SD bias 都在 0 附近波动，说明效应量标准误估计较准确。

此外，附表 7(在线补充材料 1)呈现了非平衡设计下的固定效应及其标准误估计准确性结果。与平衡设计下的结果相比，非平衡设计下的 RMSE 更大，95%CI 更宽。

表 2 研究一平衡设计效应量为 0.5 时水平 1 自变量固定效应及其标准误估计准确性

criteria	*I*	*J*									
		10	20	30	50	70	100	150	200	250	300
rbias	10	**0.008**	**0.009**	**0.001**	**0.007**	**0.001**	**0.012**	**0.001**	**0.001**	**0.001**	**0.000**
	30	**0.007**	**0.002**	**0.006**	**0.017**	**0.003**	**0.003**	**0.007**	**0.001**	**0.001**	**0.003**
	50	**0.024**	**0.009**	**0.004**	**0.011**	**0.003**	**0.002**	**0.011**	**0.005**	**0.013**	**0.002**
	70	**0.003**	**0.004**	**0.019**	**0.001**	**0.017**	**0.001**	**0.001**	**0.001**	**0.004**	**0.005**
	100	**0.019**	**0.005**	**0.004**	**0.000**	**0.004**	**0.002**	**0.004**	**0.003**	**0.002**	**0.004**
	200	**0.013**	**0.007**	**0.015**	**0.007**	**0.014**	**0.010**	**0.001**	**0.001**	**0.002**	**0.003**
	400	**0.026**	**0.025**	**0.008**	**0.003**	**0.008**	**0.004**	**0.001**	**0.003**	**0.000**	**0.002**
	600	**0.016**	**0.011**	**0.007**	**0.003**	**0.005**	**0.005**	**0.004**	**0.002**	**0.003**	**0.006**
	800	**0.005**	**0.010**	**0.010**	**0.004**	**0.001**	**0.013**	**0.003**	**0.005**	**0.000**	**0.001**
width	10	1.411	1.036	0.861	0.709	0.633	0.565	0.506	0.476	0.458	0.444
	30	1.197	0.860	0.713	0.573	0.498	0.434	0.376	0.343	0.321	0.306
	50	1.151	0.827	0.685	0.542	0.468	0.403	0.343	0.309	0.286	0.270
	70	1.125	0.817	0.669	0.530	0.453	0.389	0.328	0.293	0.269	0.252
	100	1.122	0.798	0.665	0.519	0.443	0.377	0.316	0.280	0.256	0.238
	200	1.091	0.786	0.649	0.505	0.431	0.362	0.301	0.265	0.240	0.221
	400	1.096	0.782	0.644	0.501	0.424	0.355	0.294	0.256	0.230	0.212
	600	1.086	0.778	0.643	0.497	0.422	0.353	0.290	0.254	0.227	0.209
	800	1.076	0.778	0.638	0.497	0.423	0.354	0.290	0.252	0.226	0.207
SE-SD bias	10	0.006	0.024	0.023	0.027	0.025	0.032	0.035	0.046	0.048	0.047
	30	-0.006	0.005	0.007	0.008	0.013	0.015	0.017	0.020	0.019	0.022
	50	-0.004	0.004	0.002	0.007	0.009	0.012	0.011	0.010	0.014	0.014
	70	0.004	-0.006	0.006	0.003	0.006	0.006	0.010	0.009	0.012	0.011
	100	-0.006	-0.007	0.002	0.004	0.003	0.004	0.008	0.005	0.007	0.008
	200	-0.002	0.006	0.000	0.000	0.004	0.005	0.002	0.002	0.003	0.004
	400	0.000	-0.011	0.007	0.002	0.001	0.001	0.002	0.004	0.004	0.002

600	0.000	-0.002	0.000	-0.007	-0.004	0.001	-0.002	0.001	0.002	0.003
800	-0.008	0.003	0.000	0.000	0.000	0.004	0.002	0.002	0.000	0.001

注：*J* 表示水平 1 样本量，*I* 表示水平 2 样本量，criteria 表示各评价指标。rbias 中加粗的为其值小于 0.1 的结果。

4.3.4 随机效应估计准确性结果

效应量大小基本不会影响随机效应估计准确性(附表 8 ~ 11，在线补充材料 1)。从附表 9 看出，平衡设计水平 1 自变量效应量为 0.5 情况下，σ^2估计值的 rbias 均小于 0.1，τ_{00}^2的估计准确性略优于ω_{00}^2，τ_{11}^2的估计准确性相对最低。附表 11 显示，与平衡设计下的结果相比，非平衡设计下τ_{00}^2和τ_{11}^2的估计偏差更大。

4.3.5 样本量规划建议

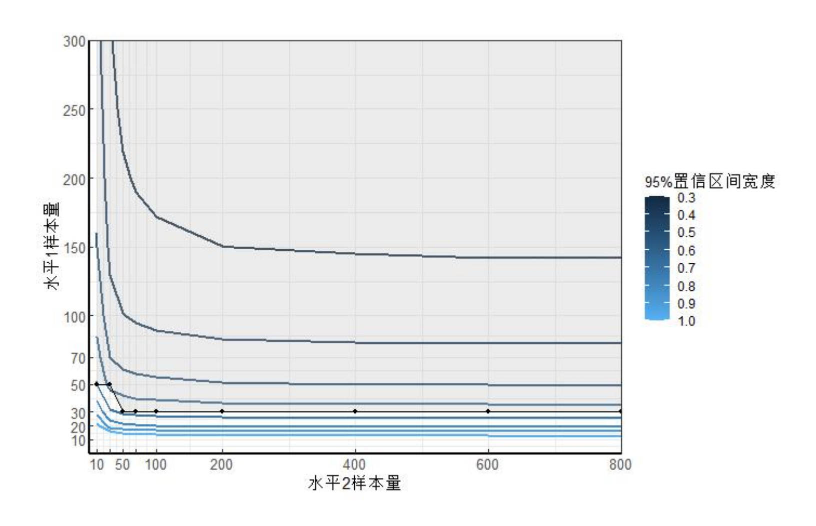
本研究提出了应用 CI 宽度等高线图给出样本量规划建议。效应量准确性主要通过 CI 宽度来反映。此外，考虑到随机效应方差也可以作为效应量指标(Hox et al., 2017)，因此也可以同时结合检验力、随机效应方差估计准确性和 CI 宽度来规划样本量。以水平 1 自变量效应量为 0.5 的情况为例，图 2(a)为检验力+CI 宽度等高线图，阴影区域表示符合检验力大于等于 0.8 标准的条件；图 2(b)为检验力+随机效应估计准确性+CI 宽度等高线图，阴影区域表示符合检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1 的条件。不同颜色对应于不同的 CI 宽度。

从图 2 可看出，首先，对于检验力，或检验力+随机效应估计准确性，两个水平样本量具有相互补偿的作用。但是，当水平 1（试次）的样本量过小时(例如，小于 30)，无论怎样增加水平 2（被试）样本量，也无法使得检验力或检验力+随机效应估计准确性达到要求。其次，95%CI 宽度受水平 1 样本量影响更大。当水平 1 样本量较小时(如 10)，即使增大水平 2 样本量，也很难减小 95%CI 宽度。最后，与图 a 相比，图 b 的阴影区域向右上移动，说明增加考虑随机效应估计准确性的要求更加严格。水平 1 自变量效应量为小、中和大情况下的等高线图见附图 1 ~ 3(在线补充材料 1)。随着效应量增大，阴影区域向下方移动，满足要求的水平 1 样本量减小。

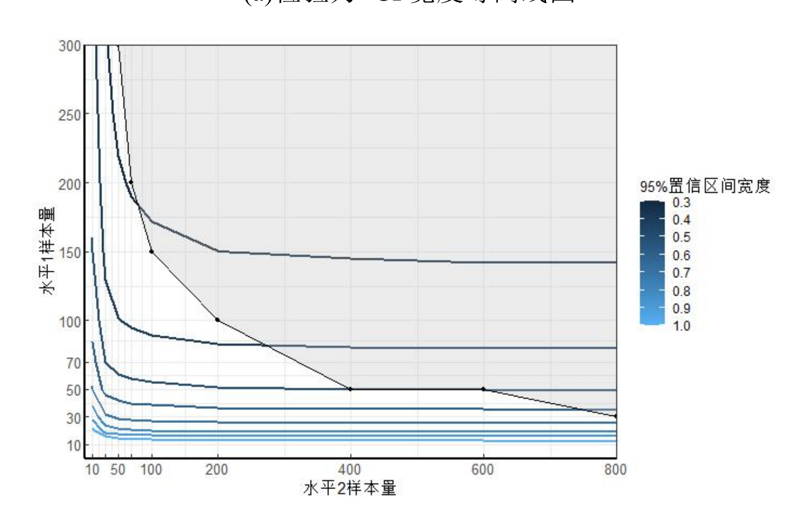
应用 CI 宽度等高线图时，首先根据阴影区域找出符合要求(检验力大于等于 0.8，或检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1)的范围。然后，在阴影区域中，通过与可接受的最宽 CI 宽度比较，得到合适的样本量组合。例如，根据图 2，满足检验力大于

等于 0.8 的标准，95% CI 宽度小于等于 0.6，则推荐水平 1 样本量=50，水平 2 样本量=30。满足检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1，95% CI 宽度小于等于 0.6，则推荐水平 1 样本量=50，水平 2 样本量=400。

从附图 3 看出，与平衡设计相比，非平衡设计下的阴影区域向上方移动，满足要求的水平 1 样本量增大，至少为 50 才能保证检验力符合要求。



(a)检验力+CI 宽度等高线图



(b)检验力+随机效应估计准确性+CI 宽度等高线图

图 2 研究一平衡设计水平 1 自变量效应量中情况下的 CI 宽度等高线图

注：图(a)中阴影区域表示符合检验力大于等于 0.8 标准的条件，图(b)中阴影区域表示符合检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1 的条件。不同 95%CI 宽度用不同颜色的等高线表示。如图例所示从 0.3 到 1.0 间隔 0.1，在图中共有 8 条依次排列的等高线。例如，0.3 对应的等高线表示线条以上的区域 95%CI 宽度在 0.3 及其以下。后同。

5 模拟研究二：基于被试变量调节效应的样本量规划研究

研究二在模型 2 的框架下，针对被试变量的调节效应(γ_{11}，跨水平交互作用)，考察 τ_{11}^2 大小和被试变量类型对模型估计结果的影响，并通过 CI 宽度等高线图提供样本量建议。

5.1 参数设置

5.1.1 固定参数设置

考虑到实际中被试变量W_i可能为分类变量(如，性别)或连续变量(如，情绪唤醒度)，研究二分为两种情境：情境 1 中，W_i为二分变量，采用偏差编码(-0.5 和 0.5)；情境 2 中，W_i为连续变量，服从标准正态分布。

与研究一类似，随机截距的固定效应γ_{00}固定为 0。研究二主要关注γ_{11}，因此，将 X_{ji} 和 W_i 的主效应固定为中等水平，即：$\gamma_{10.std} = 0.5$，$\gamma_{01.std} = 0.5$(情境 1)，$\gamma_{01.std} = 0.3$(情境 2)。为简化研究，参考检验力分析研究的普遍设计(例如, Arend & Schäfer, 2019)，将$\gamma_{11.std}$也固定为中等水平，即：$\gamma_{11.std} = 0.5$(情境 1)，$\gamma_{11.std} = 0.3$(情境 2)(Cohen, 2013)。

与研究一类似，残差方差设定为$\sigma^2 = 1$。情境 1 中，在$\tau_{11.std}^2 = 0.01$(小)，0.09(中)和 0.25(大)三种水平下(Arend & Schäfer, 2019)，根据公式(14)，可得到三种水平下的 τ_{11}^2=0.01,0.09 和 0.25。

利用τ_{11}对标准化的跨水平交互效应进行调整，得到用于产生数据的总体模型的固定效应参数(Arend & Schäfer, 2019)[7]。

$$\gamma_{11} = \gamma_{11.std} \times \tau_{11}. \tag{15}$$

[7] 在多水平模型中，$\gamma_{11.std} = \gamma_{11} * SD_{predictor} / SD_{outcome}$。当$W_i$为分类变量时，$\gamma_{11.std}$为部分标准化的回归系数，即只对因变量标准化($SD_{outcome} = \tau_{11}$，$\gamma_{11.std} = \gamma_{11}/\tau_{11}$；当$W_i$为连续变量时，由于自变量已经标准化（$SD_{predictor} = 1$），则$\gamma_{11.std} = \gamma_{11}/\tau_{11}$为完全标准化的回归系数。

因此，在情境 1 中，随机斜率方差的三种水平下γ_{11}=0.05，0.15 和 0.25；在情境 2 中，固定$\tau_{11.std}^2$为中等水平(0.09)，可得到γ_{11}=0.09。γ_{11}表示被试变量对实验效应的调节效应。在情境 1 中，γ_{11}表示$W_i=-0.5$的被试和$W_i=0.5$的被试在两个实验水平上结果差异的差异。在情境 2 中，γ_{11}表示W_i越高/越低的被试，在两个实验水平上结果的差异越大/越小。ICC 固定为中等水平。ω_{00}^2固定为 0.2。每种条件下数据重复模拟 *N*=1000 次。

5.2.2 变化参数设置

情境 1 中，在$\tau_{11.std}^2$分别为 0.01,0.09 和 0.25(公式(14))时分别进行样本量规划。同时，为考察非平衡设计对样本量规划的影响，增加被试变量两个类别样本量不等的情况(1:4)。样本量设置与研究一相同。调用 samplesize_LMEM.R 函数运行得到结果。

5.2 评价指标

与研究一相同。

5.3 研究结果

5.3.1 收敛情况

研究二中 LMEMs 的收敛率见附表 12、13(在线补充材料 1)。可以看出，当τ_{11}^2小，W_i为分类变量时，在部分条件下，收敛率低于 0.7。甚至在有些条件下(*I*=800，*J*=250 或 300)，仅有不到一半的重复收敛。**说明当τ_{11}^2较小时，采用随机斜率模型可能会带来不收敛的问题。其余各条件下基本不存在收敛问题，收敛率普遍在 0.7 以上。W_i为分类变量或连续变量、是否为平衡设计对收敛率几乎没有影响。**

5.3.2 检验力结果

各条件下γ_{11}检验力结果如附表 14、15(在线补充材料 1)所示。可以看出，**τ_{11}^2越大，检验力越大。W_i为连续变量得到的检验力普遍大于W_i为分类变量的情况，这可能与连续变量提供的信息量更多有关。**随着两个水平样本量增加，尤

其是水平 2 样本量增加，检验力增加。与研究一不同，研究二中的检验力受水平 2 样本量影响更大，这是因为研究二中的检验力是针对水平 2 自变量计算的，受被试量影响更大，而研究一中的检验力针对水平 1 自变量计算，受试次数影响更大。此外，**非平衡设计的检验力普遍小于平衡设计的结果。**

5.3.3 效应量及其标准误估计准确性结果

调节效应量及其标准误估计准确性结果见附表 16 ~ 20(在线补充材料 1)。可以看出，**τ_{11}^2 不同的条件下，bias，rbias，95%CP 和 SE-SD bias 的结果非常一致，都较小。随着 τ_{11}^2 增加，RMSE 增大，95%CI 变宽。**

与研究一不同，研究一中水平 1 自变量(实验效应)估计准确性更受水平 1 样本量影响，而研究二中跨水平交互效应估计准确性更受水平 2 样本量影响。在 W_i 为分类变量且 τ_{11}^2 为中等($\tau_{11}^2=\tau_{11.std}^2=0.09$)的情况下，根据公式(15)，计算效应量小和大条件的标准值分别为 0.06(0.2×0.3)和 0.24(0.8×0.3)。则定义可接受的最宽 95%CI 宽度为 0.24-0.06=0.18。从附表 17 看出，部分条件下 95%CI 过宽。只有当水平 2 样本量为 400，且水平 1 样本量在 50 及以上，或者水平 2 样本量在 600 及以上，且水平 1 样本量在 20 及以上时，能够满足 95%CI 宽度小于 0.18。

根据公式(15)，计算在 τ_{11}^2 小情况下，效应量小和大条件的标准值分别为 0.02(0.2×0.1)和 0.08(0.8×0.1)。τ_{11}^2 大情况下，效应量小和大条件的标准值分别为 0.1(0.2×0.5)和 0.4(0.8×0.5)。则定义两种情况下可接受的最宽 95%CI 宽度分别为 0.08-0.02=0.6 和 0.4-0.1=0.3。可以看出，τ_{11}^2 大的条件下 CI 宽度符合要求的条件多于 τ_{11}^2 小的条件。

W_i 为分类变量和连续变量得到的 bias，rbias，95%CP 和 SE-SD bias 的结果非常一致，都较小。W_i 为连续变量时得到的 RMSE 较小(见附表 18)，95%CI 较窄。根据公式(15)，效应量为小和大时 γ_{11} 分别为 0.03(0.1×0.3)和 0.15(0.5×0.1)。定义可接受的最宽 95%CI 宽度为 0.15-0.03=0.12。

此外，与平衡设计下的结果相比，非平衡设计下的 RMSE 更大，95%CI 更宽。

5.3.4 随机效应估计准确性结果

附表 21 ~ 25(在线补充材料 1)呈现了随机效应估计 rbias 结果。从表中看出，首先，与研究一类似，τ_{11}^2 **大小、W_i 类型和是否为平衡设计基本不会影响σ^2 和 ω_{00}^2 估计的准确性。**σ^2估计值的 rbias 在各样本量条件下均达到小于 0.1 的标准。其次，**当W_i 为分类变量时，随着τ_{11}^2 增加，τ_{00}^2 的估计准确性降低，τ_{11}^2 的估计准确性增加。**具体来看，当τ_{11}^2小时，几乎所有样本量条件下τ_{11}^2估计值的 rbias 都大于 0.1。进一步计算其 bias 发现，此时大部分情况下会存在高估τ_{11}^2的问题。当τ_{11}^2大时，所有样本量条件下τ_{00}^2估计值的 rbias 都大于 0.1。进一步计算其 bias 发现，此时大部分情况下存在高估τ_{00}^2的问题。最后，**当W_i 为连续变量时，τ_{00}^2的估计准确性略高于分类变量的情况。**

5.3.5 样本量规划建议

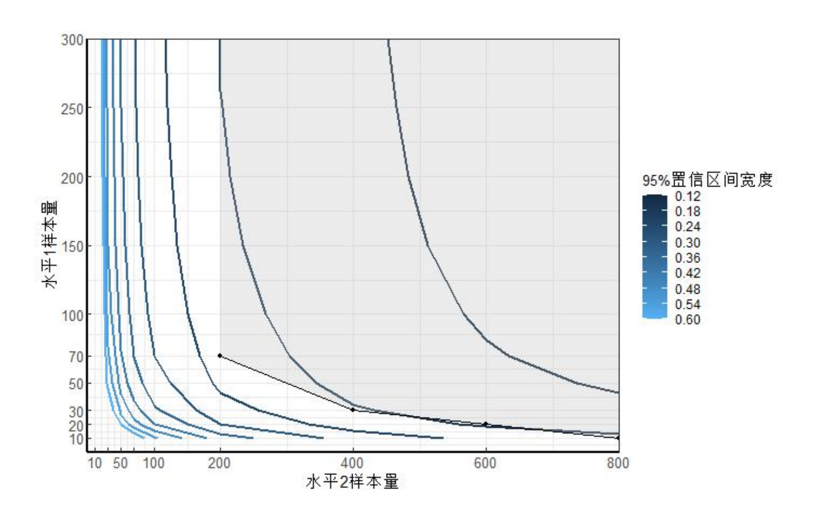
以平衡设计τ_{11}^2中等为例，图 3 和 4 分别表示了W_i 为分类变量和连续变量情况下的 CI 宽度等高线图。从图中看出，与研究一不同，95%CI 宽度受水平 2 样本量影响更大。当水平 2 样本量较小时，即使增大水平 1 样本量，也很难减小 95%CI 宽度。这可能与研究二关注的W_i 是水平 2 变量有关。此外，与研究一相比，研究二中同时满足检验力和随机效应估计准确性标准(图 b)的阴影区域相比只满足检验力大于等于 0.8 的标准(图 a)向右上移动的幅度较小，说明对于研究二来说，检验力和检验力+随机效应估计准确性标准的严格程度基本相当。并且，与研究一相比，研究二中满足检验力标准和同时满足检验力和随机效应估计准确性标准的阴影区域向右上方移动，说明对于研究二来说，需要更大的样本量组合才能保证达到要求。W_i 为分类变量和连续变量的情况下，检验力，检验力和随机效应估计准确性符合标准的区域几乎相当，W_i 为连续变量时，检验力符合标准的区域向略向下方移动，说明满足要求所需的水平 1 样本量略小。并且，W_i 为连续变量时，95%CI 更窄，同等宽度的等高线向左移动。

根据图 3，在满足检验力大于等于 0.8 的标准的情况下，如果 95%CI 宽度小于等于 0.18，

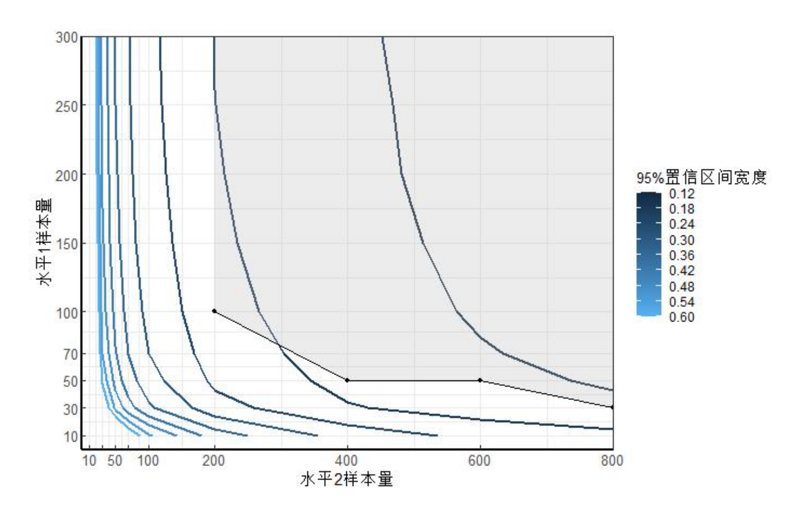
则推荐水平 1 样本量=50，水平 2 样本量=400。在满足检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1 的情况下，如果 95%CI 宽度小于等于 0.18，则推荐水平 1 样本量=50，水平 2 样本量=400。根据图 4，在满足检验力大于等于 0.8 的标准的情况下，如果 95%CI 宽度小于等于 0.12，则推荐水平 1 样本量=50，水平 2 样本量=200。在满足检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1 的情况下，如果 95%CI 宽度小于等于 0.12，则推荐水平 1 样本量=100，水平 2 样本量=200，或者水平 1 样本量=50，水平 2 样本量=400。

平衡设计W_i为分类变量情况下，τ_{11}^2小和大的 CI 宽度等高线图见附图 4 和 5(在线补充材料 1)。当τ_{11}^2小时，阴影区域向右上移动，满足要求的样本量增大；当τ_{11}^2大时，满足检验力要求的阴影区域略向下移动，满足要求的水平 1 样本量略减小，此时没有同时满足检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1 的条件。

从附图 6(在线补充材料 1)可以看出，与平衡设计相比，非平衡设计下的阴影区域向右方移动，说明满足要求的水平 2 样本量增大，至少为 400 才能保证检验力符合要求。

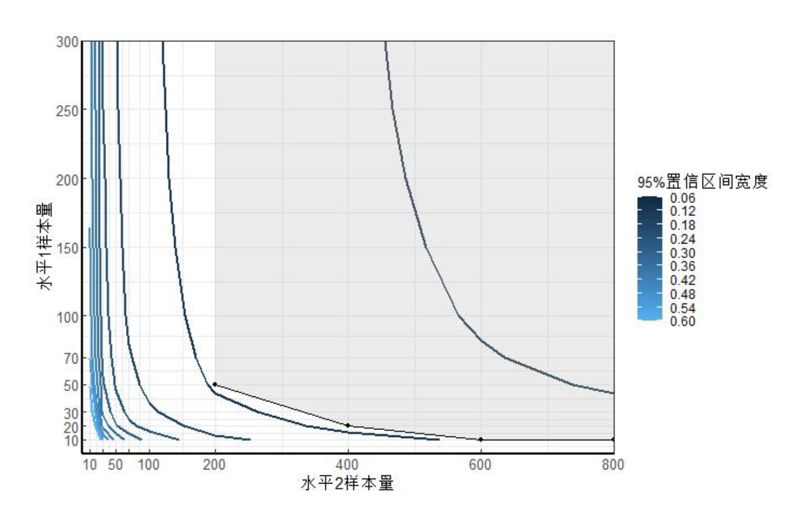


(a)检验力+CI 宽度等高线图

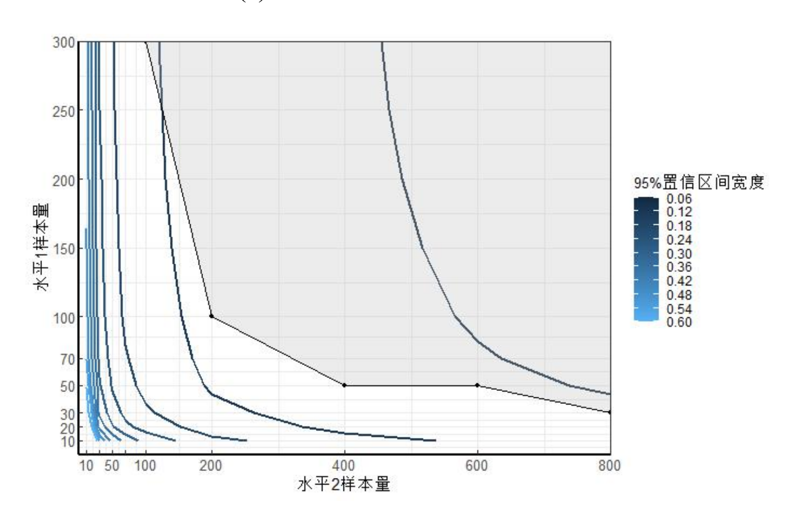


(b)检验力+随机效应估计准确性+CI 宽度等高线图

图 3 研究二平衡设计τ_{11}^2中等且W_i为分类变量时的 CI 宽度等高线图



(a)检验力+CI 宽度等高线图



(b)检验力+随机效应估计准确性+CI 宽度等高线图

图 4 研究一平衡设计τ_{11}^2中等且W_i为连续变量时的 CI 宽度等高线图

6 实例演示

本部分将通过一个例子，说明在实际中如何运用本研究开发的函数生成 CI 宽度等高线图，指导样本量规划。

假设研究者想考察某些人格特征(如诚实、道德、幽默等)是否会影响其对异性的吸引力。可参考一项关于忠诚对异性吸引力影响的类似研究(Xu et al., 2020)。该研究采用刺激不重复的单因素被试内实验设计，给被试依次呈现异性的头像，同时附上描述他们在以往恋爱关系中是否忠诚的句子，让被试对每个异性的吸引力程度等进行评分，其中忠诚与否(忠诚、不

忠诚)为被试内因素，每个条件下有 20 个不重复的刺激。研究结果显示，表现出忠诚行为的潜在伴侣的吸引力评分显著高于不忠诚的潜在伴侣。研究者可以参考本文提出的方法开展样本量规划。

首先，选取用于生成模拟数据的参数。采用借鉴前人类似研究结果设置参数。对于 Xu 等(2020)的原始数据，以是否忠诚为自变量，以面孔吸引力评分为因变量(需标准化)，将数据与本研究模型 1 拟合，估计参数。具体语句和结果请参见在线补充材料 4。根据结果，计算可得：$\gamma_{10.std}$=0.578，γ_{00}=0.000，$ICC = \tau_{00}^2/(\tau_{00}^2+\sigma^2) = 0.223$，$\tau_{11.std}^2$=0.249，$\sigma^2$=0.779，$\omega_{00}^2$=0.017。

然后，设置参数，调用函数，生成评价指标结果和 CI 等高线图。设定重复次数为 *N*=1000，水平 1 样本量包含 6 个水平：40,80,120,200,300,400。水平 2 样本量包含 6 个水平：10,30,50,70,100,200。自变量两个条件试次数相等。可接受的最宽 95%CI 宽度为 0.8-0.2=0.6。预设图中 95%CI 宽度的刻度为 kd <- c(0.3,0.4,0.5,0.6,0.7,0.8)。调用函数的语句如图 5 所示。

```
source("samplesize_LMEM.R")
N <- 1000
I <- c(10,30,50,70,100,200)
J <- c(40,80,120,200,300,400)
P1 <- 0.5
P2 <- 0.5
#input 95%CI breaks
kd <- c(0.3,0.4,0.5,0.6,0.7,0.8)
#Model1
getModelOne(I,J,P1,P2,N,0.5775,0,0.223098,0.24948,0.779,0.01706)
generatePicData("modelOne_evaluation_accuracy",kd,c(0, max(I)),c(0, max(J)),I,J,I)
```

图 5 实例演示调用函数开展样本量规划语句

最后，运行程序得到评价结果文件“modelOne_evaluation_accuracy.csv”，和检验力+CI 宽度等高线图(见图 6)[8]。根据图中所示，在满足检验力大于等于 0.8 的标准的情况下，95%CI 宽度均小于等于 0.6，则最小的推荐样本量为：被试量为 20 时，共需要 80 个试次；被试量为 30 时，共需要 60 个试次；被试量为 70 时，共需要 40 个试次。

[8] 由于本例中没有同时满足检验力大于等于 0.8 并且所有随机效应估计值 rbias 小于 0.1 的条件，因此无法生成检验力+随机效应估计准确性+CI 宽度等高线图。

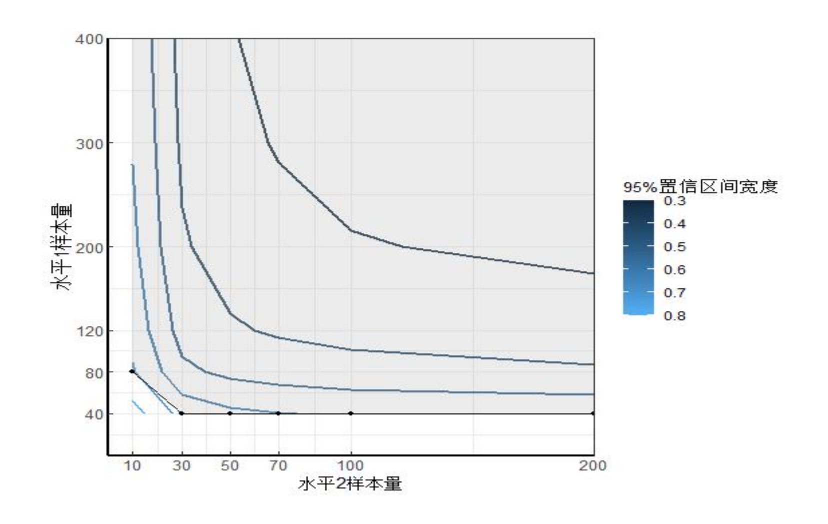


图 6 实例演示检验力+CI 宽度等高线图

7 讨论

7.1 主要研究结果

本研究针对线性混合效应模型，采用模拟方法，以被试内实验效应和被试间变量的调节效应为例，实现基于检验力与效应量准确性分析的样本量规划。并通过两个模拟研究，考察实验效应、随机斜率大小、被试变量类型和是否为平衡设计对样本量推荐结果的影响，说明 CI 宽度等高线图在样本量规划中应用。旨在为实践研究者基于具体研究实现样本量规划提供方法指导和便利工具。研究得到的主要结果如下。

第一，从收敛情况来看，对于模型 1 来说，基本不存在收敛问题。对于模型 2 来说，当随机斜率方差小时，部分条件下会存在一定程度的不收敛问题。

第二，从检验力来看，效应量越大，检验力越大。变量类型为分类变量时的检验力低于连续变量。平衡设计下的检验力普遍大于非平衡设计下的结果。此外，检验力与样本量的关系还受所考察效应所属水平的影响。水平 1 自变量效应的检验力主要受水平 1 样本量影响，水平 2 自变量效应的检验力主要受水平 2 样本量影响。两个水平的样本量具有一定程度的补偿作用，增加关注效应所在水平的样本量能更好地补偿另一水平小样本量的问题。

第三，从效应量及其标准误估计准确性来看，在拟合模型定义准确的情况下，固定效应点估计值都较准确。但是，其 CI 宽度会受到是否为平衡设计和随机效应的影响。非平衡设计下的 CI 普遍更宽。对于水平 2 变量的调节效应，随机斜率方差越大，CI 越宽，效应量估计的标准误越大。各条件下效应量估计标准误的准确性都较高。

第四，从随机效应估计准确性来看，残差方差估计准确性都较高。随机截距和随机斜率

方差估计准确性会受是否为平衡设计和随机斜率方差大小的影响。对于仅含被试内自变量的模型，非平衡设计下随机截距方差和随机斜率方差的估计准确性更低。随机斜率方差越大，随机截距方差的估计准确性越低，随机斜率方差的估计准确性越高。随机斜率方差小时，会高估随机斜率方差，随机斜率方差大时，会高估随机截距方差。

7.2 实践建议

本研究期望以两种较典型的线性混合效应模型为例，说明规划样本量的方法。基于研究过程和结果，提出以下建议。

首先，样本量规划需同时结合检验力与效应量准确性分析结果。传统的样本量规划主要基于检验力分析展开(例如, Schultzberg & Muthén, 2018)，确保推荐样本量能够满足检验力要求(0.8 及以上)。但是，随着目前越来越多的学术期刊和研究机构呼吁在报告显著性的基础上，报告效应量及其CI，对效应量估计准确性的要求也日益受到重视(Maxwell et al., 2008)。其实，基于检验力与基于 CI 宽度规划样本量既有联系，又有区别。两种方法的联系在于，无论是基于检验力还是 CI 宽度规划样本量，都与效应量的标准误有关。在固定效应模型下，CI 可以定义为$[T-1.96SE, T+1.96SE]$(T 表示效应量估计值，SE 表示标准误)。在随机效应模型下，随机效应的方差部分会加入到标准误的计算中，因此，与固定效应模型相比，会得到更大的标准误(SE^*)，此时效应量的 CI $[T-1.96SE^*, T+1.96SE^*]$会更宽。无论是固定效应模型还是随机效应模型，效应量的标准误越小，效应量估计值的 CI 就越窄，效应量的估计值就越准确。在假设效应量不为0的情况下，越窄的CI就越不可能包括0，会得到更大的检验力(Cohn & Becker, 2003)。两种方法的区别在于，真实的总体效应量越大，其 CI 就越不可能包括 0，因此检验力越大；但 CI 宽度不受影响。因此，效应量越大，基于检验力规划的样本量越小，而基于 CI 宽度规划的样本量不变，这也与本研究结果一致。本研究发现，基于检验力分析与效应量估计准确性推荐的样本量不一定相等。例如，从研究一的图 2(b)中发现，在水平 1 自变量效应量中等的情况下，当水平 2 样本量为 50 时，水平 1 只需要 30 个试次，就能保证检验力大于 0.8。但此时实验效应的效应量估计值 CI 宽度为 0.7 左右，大于可接受的最宽 CI 宽度。因此，应当同时结合二者结果确定推荐的样本量。

其次，在基于模拟方法进行样本量规划时，应当谨慎确定产生数据模型的参数。通过检验力与效应量准确性分析开展样本量规划需要研究者预先设定一些模型参数(如预期效应量，ICC 等)，以便基于特定模型产生数据。特别说明的是，本研究主要目的是说明样本量

规划的方法及 CI 等高线图的使用，参数设置不一定代表实际中的大多数情况。在实际研究中，研究者可以从前人已发表的类似研究，自己的预研究，相关主题的元分析结果，或者由同领域专家确定最小的重要效应来获得这些参数值(Pek & Park, 2019)。然而，也有研究者指出，这种直接使用效应量点估计值代替其真值(预期效应量)的方式忽略了其不确定性(uncertainty with regard to the unknown population effect size, Pek & Park, 2019)，会得到有偏差的结果。因此，一些研究者提倡使用考虑了不确定性问题的方法(如贝叶斯混合方法，Pek & Park, 2019)开展样本量规划。

然后，实践研究者可以根据具体研究需要，结合本研究提出的两种 CI 宽度等高线图确定推荐的样本量。本研究参考 Baker 等(2021)检验力等高线图的思路，提出 CI 宽度等高线图，能够便于研究者同时参考多种要求，找到最合适的样本量。研究者可根据实际研究对结果的要求，确定选用某种 CI 宽度等高线图。如果研究者仅关注检验力和效应量估计的准确性，可选用检验力+CI 宽度等高线图。如果研究者在此基础上，还关注随机效应估计的准确性，以便对个体差异的原因进行进一步分析(如应用混合效应均值——方差模型，Williams et al., 2021)，或者进一步准确计算包含随机效应的R^2指标(例如, Rights & Sterba, 2019)，可以选用检验力+随机效应估计准确性+CI 宽度等高线图。对于 CI 宽度，研究者可以参考本研究的做法，也可以参考前人研究中效应量的 CI 宽度，或结合自己研究中效应量报告精度的需要确定临界值。

最后，在实际研究中，样本量规划是结合检验力、效应量准确性与研究成本等的综合考虑。如果仅考虑检验力和效应量准确性，往往会导致规划的样本量很大。较大的样本量会带来研究成本的显著增加。尤其是对于一些人力、物力成本较大的研究(例如，应用功能性磁共振成像的研究等)，大幅增加被试量往往不现实。因此，一些研究者提出了结合研究成本函数综合得到推荐样本量的方法(例如, Baker et al., 2021)，以保证样本量既能够满足检验力等要求，又使得研究成本尽可能最小。例如，在 Baker 等(2021)开发的网页中，就结合了每名被试的成本，计算推荐样本量。该网页中得到的推荐样本量是检验力达到 80%且$I \times (J + 成本)$最小的点。除了研究成本，在实际中不同研究确定样本量会结合不同的研究限制，并有优先考虑的要求顺序等级。应用研究者可结合具体研究需求，在本研究提供的方法基础上开展样本量规划。

7.3 未来研究展望

本研究具有一定的局限性, 未来研究可以从三个方面加以改进。首先，本研究的模拟研

究只考察了实验效应、随机斜率大小、被试变量类型、是否为平衡设计的影响，很多因素设置为固定水平。未来研究可考察随机截距和随机斜率的协方差，刺激的随机效应方差等因素对检验力和效应量准确性的影响，获得更加丰富的结果。其次，本研究以刺激嵌套于实验条件，并且刺激和实验效应没有交互的被试内实验设计为例探讨样本量规划的问题，并假设实验条件是含有两个类别的分类变量，因变量是连续变量。未来研究可以拓展到其他类型的实验设计，或者自变量为连续变量，因变量为分类变量等情境，探讨基于线性混合效应模型的样本量规划，丰富函数功能。最后，本研究没有考虑预期效应量的不确定性问题，不能反映实践中研究设计面临的现实困境。未来研究可以借鉴 Pek 和 Park(2019, 2023)的思路，通过检验力和效应量准确性的分布实现样本量规划。

参考文献

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on monte carlo simulation. *Psychological Methods*, *24*(1), 1−19.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295−314.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Grothendieck, G. (2011). Package 'lme4'. *Linear mixed-effects models using S4 classes. R package version*.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389−411.

Cho, S. J., De Boeck, P., & Lee, W. Y. (2017). Evaluating testing, profile likelihood confidence interval estimation, and model comparisons for item covariate effects in linear logistic test models. *Applied Psychological Measurement*, *41*(5), 353−371.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*(3), 243−253.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7* (4), 493–498.

Hecht, M., & Zitzmann, S. (2021). Sample size recommendations for continuous-time models: Compensating shorter time series with larger numbers of persons and vice versa. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 229–236.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and*

applications. Routledge.

Jiang, Y., Jiang, C., Hu, T., & Sun, H. (2022). Effects of emotion on intertemporal decision-making: Explanation from the single dimension priority model. *Acta Psychologica Sinica*, *54*(2), 122−140.

[蒋元萍, 江程铭, 胡天翊, 孙红月. (2022). 情绪对跨期决策的影响: 来自单维占优模型的解释. *心理学报*, *54*(2), 122−140.]

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625.

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226−243.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363−385.

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology*, *5*, 311. https://doi.org/10.3389/fpsyg.2014.00311

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528−2543.

Lee, W. Y. (2018). *Generalized linear mixed effect models with crossed random effects for experimental designs having non-repeated items: Model specification and selection* (Unpublished Doctoral dissertation). Vanderbilt University.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537−563.

Murayama, K., Usami, S., & Sakaki, M. (2022). Summary-statistics-based power analysis: A new and practical method to determine sample size for mixed-effects modeling. *Psychological Methods*. *27*(6), 1014−1038

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Almenberg, A. D., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science.

Annual Review of Psychology, *73*, 719−748.

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, *24*(5), 590−605.

Park, J., & Pek, J. (2023). Conducting Bayesian-classical hybrid power analysis with R package hybridpower. *Multivariate Behavioral Research*, *58*(3), 543−559.

R Development Core Team. (2019). R: *A language and environment for statistical computing* [Computer software Manual]. Vienna, Austria. http://www.Rproject.org (ISBN 3-900051-07-0).

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309−338.

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 495−515.

Usami, S. (2020). Confidence interval‐based sample size determination formulas and some mathematical properties for hierarchical data. *British Journal of Mathematical and Statistical Psychology*, *73*, 1−31.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2021). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*, *26*(1), 74−89.

Xu, L., Becker, B., Luo, R., Zheng, X., Zhao, W., Zhang, Q., & Kendrick, K. M. (2020). Oxytocin amplifies sex differences in human mate choice. *Psychoneuroendocrinology*, *112*, 104483. https://doi.org/10.1016/j.psyneuen.2019.104483

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, *46*(4), 1184−1198.

Confidence Interval Width Contours: Sample Size Planning for Linear Mixed-Effects Models

LIU Yue[1], XU Lei[1], LIU Hongyun[2,3], HAN Yuting[4], YOU Xiaofeng[5], WAN Zhilin[1]

([1] *Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu* 610066, *China*)

([2] *Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875, *China*)

([3] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*) ([4] *School of Psychology, Beijing Language and Culture University, Beijing*, 100083, *China*) ([5] *School of Mathematics and Information Science, Nanchang Normal University, Nanchang* 360111, *China*)

Abstract

Hierarchical data, which is observed frequently in psychological experiments, is usually analyzed with the linear mixed-effects models (LMEMs), as it can account for multiple sources of random effects due to participants, items, and/or predictors simultaneously. However, it is still unclear of how to determine the sample size and number of trials in LMEMs. In history, sample size planning was conducted based purely on power analysis. Later, the influential article of Maxwell et al. (2008) has made clear that sample size planning should consider statistical power and accuracy in parameter estimation (AIPE) simultaneously. In this paper, we derive a confidence interval width contours plot with the codes to generate it, providing power and AIPE information simultaneously. With this plot, sample size requirements in LMEMs based on power and AIPE criteria can be decided. We also demonstrated how to run sensitivity analysis to assess the impact of the magnitude of experiment effect size and the magnitude of random slope variance on statistical power, AIPE and the results of sample size planning.

There were two sets of sensitivity analysis based on different LMEMs. Sensitivity analysis Ⅰ investigated how the experiment effect size influenced power, AIPE and the requirement of sample size for within-subject experiment design, while sensitivity analysis Ⅱ investigated the impact of random slope variance on optimal sample size based on power and AIPE analysis for the cross-level interaction effect. The results for binary and continuous between-subject variables were compared. In these sensitivity analysis, two factors regarding sample size varied: number of subjects (I=10, 30, 50, 70, 100, 200, 400, 600, 800), number of trials (J=10, 20, 30, 50, 70, 100, 150, 200, 250, 300). The additional manipulated factor was the effect size of experiment effect (standard coefficient of experiment condition= 0.2, 0.5, 0.8, in sensitivity analysis Ⅰ) and the magnitude of random slope variance (0.01, 0.09 and 0.25, in sensitivity analysis Ⅱ). A random

slope model was used in sensitivity analysis Ⅰ, while a random slope model with level-2 independent variable was used in sensitivity analysis Ⅱ. Data-generating model and fitted model were the same. Estimation performance was evaluated in terms of convergence rate, power, AIPE for the fixed effect, AIPE for the standard error of the fixed effect, and AIPE for the random effect.

The results are as following. First, there were no convergence problems under all the conditions , except that when the variance of random slope was small and a maximal model was used to fit the data. Second, power increased as sample size, number of trials or effect size increased. However, the number of trials played a key role for the power of within-subject effect, while sample size was more important for the power of cross-level effect. Power was larger for continuous between-subject variable than for binary between-subject variable. Third, although the fixed effect was accurately estimated under all the simulation conditions, the width 95% confidence interval (95%width) was extremely large under some conditions. Lastly, AIPE for the random effect increased as sample size and/or number of trials increased. The variance of residual was estimated accurately. As the variance of random slope increased, the accuracy of the estimates of variances of random intercept decreased, and the accuracy of the estimates of random slope increased.

In conclusion, if sample size planning was conducted solely based on power analysis, the chosen sample size might not be large enough to obtain accurate estimates of effects size. Therefore, the rational for considering statistical power and AIPE during sample size planning was adopted. To shed light on this issue, this article provided a standard procedure based on a confidence interval width contours plot to recommend sample size and number of trials for using LMEMs. This plot visualizes the combined effect of sample size and number of trials per participant on 95% width, power and AIPE for random effects. Based on this tool and other empirical considerations, practitioners can make informed choices about how many participants to test, and how many trials to test each one for.

Key words linear mixed-effects models, multilevel models, power analysis, effect size, confidence interval width